Universidad de los Andes Facultad de Ingeniería	Ingeniería de Sistemas y Computación **Maestría** ISIS4517 – Análisis de Big Data – Sección 01 http://sistemas.uniandes.edu.co/~isis4517 Semestre: 2014-10 Horario: Martes – 18:00 a 20:50	Accredited ABET Engineering Accreditation Commission

Taller 3 – Repositorios NoSQL y análisis básico de contenido

Objetivo

- Utilizar el entorno MongoDB en la construcción de soluciones altamente escalables para la el procesamiento de información.
- Utilizar un repositorio de datos NoSQL para la consulta de documentos sobre dataset reales.
- Utilizar herramientas de análisis de polaridad en textos.
- Experimentar con infraestructuras que permiten la escalabilidad de procesamiento a través de la paralelización de procesos

Prerrequisitos

- Herramientas y lenguajes para desarrollo de aplicaciones Web. Por ejemplo, Java, JSP, Phyton, etc.
- Conocimiento básico de Unix y ambientes de virtualización
- Conocimiento básico de MongoDB
- Conocimiento básico de técnicas de modelaje de contenido de documentos.

Metodología

- Se trabaja de acuerdo con los lineamientos generales del curso.
- Se realiza una entrega por grupo
- Utilice para el documento las pautas de elaboración de documentos técnicos que encuentra en Sicua+.

Enunciado

1. **Análisis de polaridad sobre Twitter**
   Visualizar la polaridad, localización y tendencias en un dataset tomado de Twitter.

   - Revise el *dataset* de tuits, estudie el formato de los datos, contenido y alcance.
   - Guarde todos los datos en MongoDb. La base de datos DEBE TENER como prefijo <GrupoNN> donde NN es el identificador de su grupo.
   - Construya consultas básicas que le permitan evaluar el alcance de MongoDB en su expresividad, forma de procesar los datos y entrega de resultados. Elabore al menos 3 consultas complejas.
   - Realice un análisis de polaridad de los tuits. Elabore un análisis de sentimientos en por lo menos 3 niveles (positivo, negativo, neutro). Sería preferible hacer uno en mayores niveles, para ello busque las herramientas adecuadas.
   - Compare su resultado de análisis de polaridad con el producido por los evaluadores en el dataset. Tenga en cuenta que no siempre se tienen resultados completos de los evaluadores.
   - Realice un análisis de frecuencia de los usuarios y los *hashtags* encontrados
   - Realice un análisis de geocodificación para los tuits que mencionan ciudades.
   - Construya una aplicación Web que permita:
     - ✓ Realizar las consultas y visualizar los resultados de forma dinámica.
     - ✓ Visualizar los resultados de análisis de polaridad comparados con los encontrados en el dataset.
     - ✓ En UNA página de visualización muestre un mapa de los tuits geocodificados (sitio y cantidad de tuits), un tagCloud con los *hashtags* más frecuentes y una visualización de la polaridad que permita ver el contenido de los tuits relacionados.
     - ✓ El sitio Web debe incluir explícitamente la citación de la fuente del dataset, tal como está especificado en http://www.infochimps.com/datasets/twitter-sentiment-dataset-2008-debates

BONO

- Tome una muestra de tuits producidos por tuiteros colombianos durante 10 días y realice el mismo ejercicio.
  - Tenga en cuenta que Twitter tiene restricciones en la cantidad de información que permite descargar diariamente.
  - Pueden unir *datasets* generados por varios grupos de forma que constituyan un *dataset* más significante. Deben eliminar elementos repetidos.

RESTRICCIONES

- EN NINGÚN CASO debe hacerse sobre la versión en línea de Twitter.
- No intente seguir los links que encuentra en los tuits.
- DEBE utilizar estrategias escalables en la solución del problema.
- DEBE utilizar MongoDB como repositorio de la información (tanto los fuentes como los resultados).
- Debe realizar la visualización de forma dinámica sobre los resultados obtenidos y almacenados.

2. **Escalabilidad en MongoDB**:
   Realice el proceso del paso 1 en escenarios diversos de escalabilidad y documente sus resultados de experimentación.

   Una vez tenga claro y funcional el proceso solicitado en el punto 1, procese los datos en diferentes escenarios de escalabilidad. Para cada escenario documente el tiempo que tardan los procesos de *consulta* y análisis.
   a. Utilice una instalación propia de MongoDB standalone en su máquina virtual que descargó al inicio del curso. NO lo haga en la máquina de publicación de resultados.
   b. Utilice un *cluster* MongoDB que estará disponible para la experimentación. Revise y documente la configuración que encuentra allí para MongoDB.

RESTRICCIONES

- Para realizar los procesos deben reservar turnos para el uso del *cluster*, de manera que sólo un grupo esté procesando al tiempo. Se sugiere programar la ejecución de los Jobs de tal manera que puedan arrancar y dejar los datos sin que requieran intervención personal.
- Los turnos serán dispuestos en horarios nocturnos, cuando baja la carga de los medios de almacenamiento y carga de máquinas virtuales en el *datacenter*. Se espera una distribución equitativa de los turnos a los grupos, aunque se programarán en libre demanda. Se utilizará una herramienta Web para la asignación de turnos.
  - El doodle de turnos será publicado en Sicua+
  - Las credenciales para el acceso al cluster serán publicadas en Sicua+

Entregable

- Muestre los resultados solicitados en una aplicación Web sencilla, que ofrezca las ***funcionalidades*** solicitadas. No olvide relacionar el número de grupo y sus integrantes en la página Web de resultados.
- Elabore un documento de **máximo 4 páginas** en el cual relacione:
  - Forma de acceso a la aplicación Web resultante.
  - Resultados del procesamiento realizado a través de la aplicación Web dispoibles e identificables fácilmente en MongoDB
  - **Métodos y tecnología** concretos utilizados en cada uno de los retos propuestos
  - **Estrategia** que le utiliza en la solución
  - El **algoritmo básico** para resolver cada uno de los retos, de manera que puedan percibirse los elementos interesantes para poner en valor en la solución
  - **Análisis de resultados obtenido.**
- **Si realiza el bono, el dataset obtenido debe ser puesto a disposición del laboratorio CODICE.**

Aspectos que el grupo decide

1. Herramientas para visualización Web
2. Lenguaje y ambiente de desarrollo

Requerimientos técnicos

1. La interacción con el usuario debe ser en una aplicación Web gráfica, sencilla pero intuitiva y bien presentada.

2. Desarrolle y despliegue la aplicación solicitada en el ambiente UNIX provisto en el curso. En particular, la aplicación de demostración debe correr en la máquina Web prevista para publicación de resultados utilizando el cluster MongoDB.

Asignación de fuentes de datos

Datos disponibles para descarga en http://www.infochimps.com/datasets/twitter-sentiment-dataset-2008-debates

Evaluación

La evaluación se hace así:

Entregable	Fecha y hora límite de entrega	Porcentaje en la evaluación
Proyecto de software desarrollado, aplicación funcional	Jueves 10 de abril 9 am	70%
Informe, sustentación y demostración	Jueves 10 de abril 9 am	30%

El cumplimiento de las restricciones técnicas es parte integral de los dos entregables. No satisfacerlos invalida TODOS LOS entregables.

Se espera que cada miembro del grupo haga una contribución igualmente significante al desarrollo de esta actividad y a las tareas definidas al interior del grupo. El trabajo por debajo de este rango tiene una penalización proporcional sobre la evaluación global de la tarea

Los resultados serán sustentados en sesión de 10 minutos por grupo en horario definido en Sicua+.

Entregables

.

Archivo de la entrega: <Taller3_NN_login1_login2_login3>.zip.

Donde NN es el número del grupo y login1 y login2 son los correspondientes a los miembros del grupo en Uniandes.

Contenido: Archivo zip con el proyecto de software y archivo .pdf con el informe de análisis. Nombre del archivo de análisis: Taller3_NN_login1_login2_login3.pdf

El no seguimiento del formato de entrega del taller tiene una penalización de **0.5/5.0** en la nota final. La no presentación a la sustentación de los resultados produce una nota final en la tarea de 0.0/5.0. El grupo COMPLETO tiene UNA oportunidad de sustentación.